ПРОГНОЗИРОВАНИЕ INTERNET-ТРАФИКА С ИСПОЛЬЗОВАНИЕМ НЕЙРОСЕТЕВОГО ПОДХОДА

Хмелевой С.В.
Кафедра АСУ, ДонНТУ.
hmelevoy_sergey@ukr.net

Abstract

Khmilovyy S.V. Forecasting of internet-traffic with use neural network approach. *An Article is devoted to a problem of forecasting of the data traffic, and also critical situations definition in channels of internet service provider (ISP). The technique of preliminary data processing and realizations of forecasting with neural networks is described.*

Актуальность

Данная работа посвящена применению и уточнению нейросетевой методики прогнозирования временных рядов применительно к задаче прогнозирования internet-трафика, а также классификации и обнаружения критических ситуаций в каналах данных Internet Service Provider (ISP). Ранее применение данной методики к таким задачам не исследовалось. Рассматриваются вопросы Data Mining, структуры нейронных сетей, методики прогнозирования временных рядов.

Данная методика приведена в [1]. Рекомендации по предварительной обработке данных приведены в [2], [3], [4]. Структура применяемой НС обсуждается в [5], [6], [7], [8].

Постановка задачи

Каждый провайдер имеет входящие и исходящие каналы передачи данных. В зависимости от топологии коммутации потоков данных на оборудовании ISP, один и тот же поток может быть входящим в одном месте и исходящим в другом, к тому же он может мультиплексироваться (становиться частью более крупного потока данных, объединяющего несколько логических потоков в одном физическом).

На рис. 1 показаны зависимости данных двух результирующих потоков, которые являются главным численным выражением загруженности каналов провайдера, от времени. Каналы In (общий входной поток, показан на рисунке зеленым) и Out (общий выходной поток, показан фиолетовым) уже включают в себе все входящие и

исходящие потоки данных ISP с учетом их мультиплексирования, объединяют несколько логических потоков в одном физическом. Эти данные, а также некоторые другие хранятся в базе данных ISP.

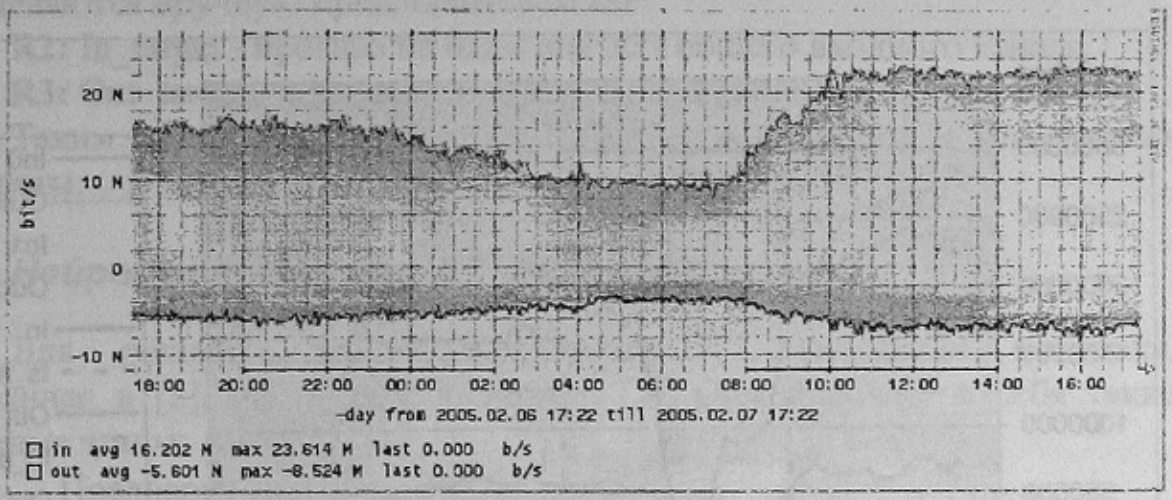

Рисунок 1 - Общие входной и выходной каналы ISP

Во время функционирования ISP время от времени происходят т.н. критические ситуации. Это резкие неожидаемые всплески или падения объема данных, проходящих через каналы In или Out, происходящие как по вине самого ISP, так и по причинам, не зависящим от провайдера. ISP необходимо вовремя определять эти ситуации, а также прогнозировать загруженность каналов (объем данных, проходящих через каналы в единицу времени) в последующие временные интервалы.

Целью данной работы является прогнозирование критических ситуаций в каналах провайдера, что поможет уменьшить убытки предприятия, связанные с перебоями в internet-трафике.

Входные данные обучающей выборки

На рис. 2 показан пример ситуации с отображением влияющих факторов (показаны разными цветами сплошной линией), аппроксимированные средние по каналам In и Out (показаны точками), а также участка, который можно считать критической ситуацией (коричневая прерывистая линия).

Основываясь на этой информации были получены следующие влияющие на In и Out факторы:

V1, V2: In0, In1 – информация по первичным входным каналам 0 и 1,

V3, V4: Out0, Out1 – информация по выходным каналам 0 и 1.

V5: In – информация по результирующему входному каналу с учетом его мультиплексирования. Ряд In рассчитывается как: In=In0 – Out1

V6: Out – информация по результирующему выходному каналу с учетом его мультиплексирования. Рассчитывается как: Out=Out0 – In1.

V7, V8, V9: In_lag1, In_lag5, In_lag10 – 1, 5 и 10 лаги ряда In,
V10, V11, V12: Out_lag1, Out_lag5, Out_lag10 – 1, 5 и 10 лаги Out.

Таким образом, глубина временного ряда обучающей выборки ограничивалась 11 точками.

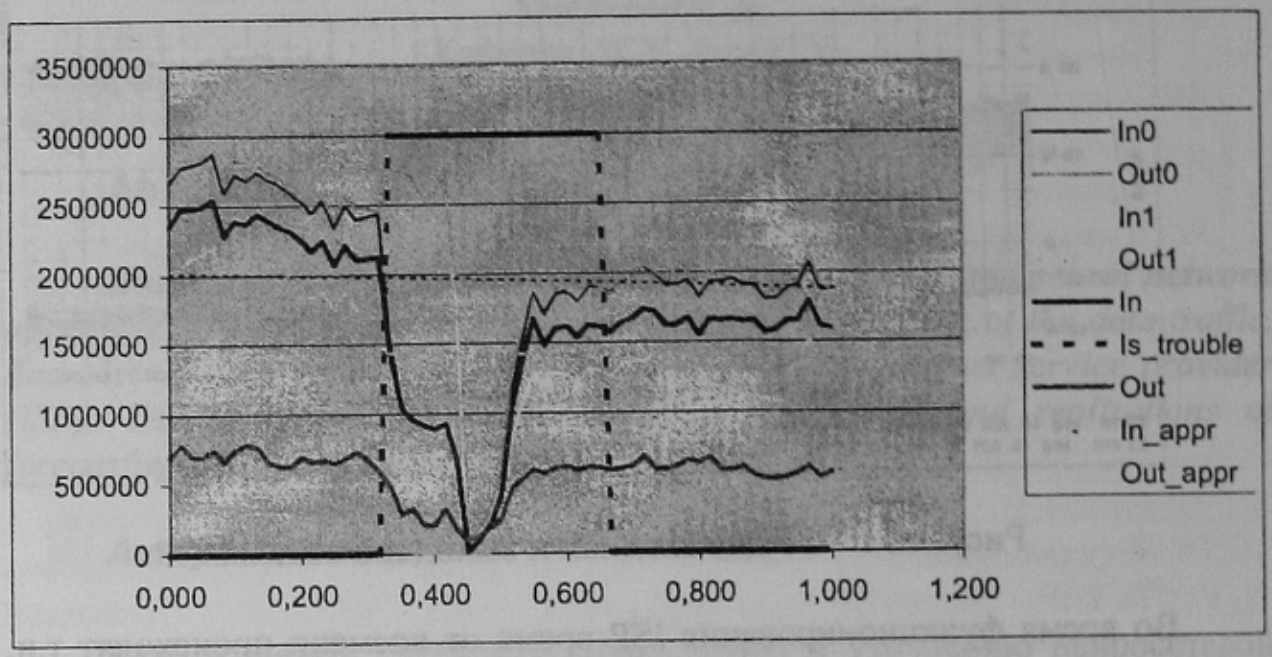

Рисунок 2 - Пример входной информации

Поскольку временные ряды входящих и исходящих каналов стохастичны, для работы необходимо использовать более сглаженные значения. Поэтому добавлены еще данные по средним последних 10 лагов общих входных и выходных каналов, а также отклонения от этих средних:

V13, V14: In_avg, Out_avg – среднее последних 10 лагов общего входного канала и выходного каналов.

V15, V16: In_delta, Out_delta – отклонение значения общих входного и выходного канала от их средних значений.

Также в процессе прогнозирования использовались расчетные сглаженные значения общих входных и выходных каналов. Для этой цели использовалась полиномиальная аппроксимация с приведением конечных значений ряда за день к начальным. Строилось 3 группы рядов – рабочие дни, субботы и воскресенья:

V17, V18: In_appr, Out_appr – расчетные значения общих входного и выходного каналов.

Остальные используемые факторы:

V19: Hours_modif – время в формате HH.MM.

V20: No_situation – номер ситуации.

Выходные данные обучающей выборки

R1: Is_trouble – наличие критической ситуации. Границы ситуаций определяются вручную представителем ISP.

R2: In_target - прогноз на один шаг для общего входного канала,

R3: Out_target – прогноз для общего выходного канала.

Таким образом, задача имеет 20 входных факторов и 3 выходных величины.

Нейросетевой подход к прогнозированию.

Для решения задачи использовалась следующая методика, показанная в [2], [4], и дополненная в [1]. Она включает в себя такие основные этапы:

- Предварительная обработка данных
- Определение параметров нейронной сети
- Обучение нейронной сети и получение результатов

В соответствии с этой схемой построена работа с сетью.

Предварительная обработка данных

Все значения факторов (V1-V18, V20) приводились к диапазону [0,1]. Фактор V19: Hours_modif преобразован к непрерывному виду [3]:

• Поскольку в 7 часов утра загрузка всех каналов минимальна, то часы (HH) сдвинуты на 7 часов. 0 данного ряда соответствует 7 часам утра, что позволяет связывать арифметическую величину часа с значениями рядов.

• Для корректного представления минут (MM) в плоскости реальных чисел они переводятся из шестидесятых долей часа в сотые доли часа.

Определение параметров нейронной сети

Согласно [1], [6], [7], [8], в качестве базовой сети была выбрана MLP-сеть с 1 скрытым и 1 выходным слоем. Количество нейронов в выходном слое равняется количеству выходов – 3. Определение глубины исторической выборки приведено ниже. Поскольку для данной предметной области есть 2 подзадачи: классификации и прогнозирования, то необходимы и 2 оценки точности прогноза. О точности задачи классификации можно судить по отношению количества ошибок к общему числу точек в выборке. О точности прогнозирования возможно судить по величине ошибки, нормированной по разнице между максимальным и минимальным значением ряда:

$$E = \frac{1}{n}\sum_{i=1}^{n} \left| \tilde{x}_i - x_i \right| / (x_{i\max} - x_{i\min}) \qquad (1)$$

Такая мера обладает одинаковой чувствительностью к ошибкам прогноза для разных значений прогнозируемой характеристики.

В качестве инструментальной среды использовался пакет Mathlab6.0 (neural network toolbox). Количество нейронов промежуточного слоя первоначально было завышенным (равным количеству точек в обучающей выборке), а затем уменьшалось, пока величина погрешности не начала увеличиваться. Активационная функция скрытого слоя – гиперболический тангенс, выходного слоя – для задачи прогнозирования линейная, для задачи классификации пороговая. Выбор метода обучения приведен ниже.

Обучение нейронной сети и получение результатов

В качестве ***метода обучения*** НС выбрана функция trainrp, которая имеет как достаточно высокую точность результата, так и достаточно малое время выполнения при предпочтительной минимизации времени выполнения. Из предоставленных ISP ситуаций данные были обработаны и составлены ***обучающая (ОВ) и тестовая (ТВ) выборки***. Распределение производилось такими способами:

1. Часть ситуаций относится к ОВ, часть – к ТВ. Данные из ситуаций располагаются в выборке последовательно.
2. Часть ситуаций относится к обучающей, часть – к тестовой выборке. Данные из ситуаций в выборке перемешаны.
3. Все данные перемешаны, и потом распределены между выборками. В ОВ и ТВ есть близкие данные из одной ситуации.

Результаты исследования приведены в таблице 1.

Таблица 1. Зависимость точности результата от вида выборки.

	Все данные перемешаны	Группировка по порядку следования	Данные перемешаны в пределах выборки	Группировка по порядку следования, короткий ряд	Данные в выборке перемешаны, короткий ряд
Эпох	800	800	800	800	800
Время на обучение	146.58	159.27	146.25	126,90	128,31
Количество данных ОВ	1729	1757	1757	1757	1757
Процент ошибки классификации ОВ	3.62	2.30	2.37	2,80	2,57
Значение функции ошибки прогнозиров. канала In ОВ	2.76	2.41	2.44	2,38	2,31
Значение функции ошибки прогнозиров. канала Out ОВ	1.57	1.52	1.64	1,39	1,38
Количество данных ТВ	733	711	711	711	711
роцент ошибки класс. ТВ	4.50	6.69	8.69	6,54	8,44
Значение функции ошибки прогнозиров. канала In ТВ	2.93	5.65	6.54	5,79	5,66
Значение функции ошибки прогнозиров. канала Out ТВ	1.71	5.59	5.52	5,58	4,69

Было произведено ***исследование оптимального времени обучения*** начиная со 100 эпох, далее количество эпох удваивается до 1600, когда происходит переобучение сети. Соответственно, оптимальное время обучения принято равным 800 эпохам.

При исследование необходимой длины ряда сокращение длины ряда до 1 члена (из обучающей выборки убрали факторы V8: In_lag5, V9: In_lag10, V11: Out_lag5, V12: Out_lag10) дало очень интересные результаты (таблица 1). Точность и прогнозирования, и классификации возросла практически на всех задачах и на всех выборках. Соответственно, особенностью данной задачи есть очень малая необходимая длина временного ряда, равная 2 членам!

Процент погрешности прогнозирования. Функция ошибки, используемая при обучении НС для задачи прогнозирования (1) не является процентом ошибки в буквальном значении этого слова. Расчет и обучение сети по такой оценочной функции как процент ошибки затруднительно, в связи с тем, что в выборке существуют значения, близкие к нулю. Для оценки действительной точности прогнозирования, необходимо привести и ориентировочные значения этого показателя. Значению используемой функции ошибки = 2.3 – 2.5 соответствует 6-7% ошибки. Таким образом, ошибка прогнозирования является достаточно значительной, хотя ошибка классификации и достаточно низка для практического использования.

Данные результаты можно объяснить следующим образом:

1. В обучающей выборке недостаточное количество влияющих факторов. Необходим поиск дополнительных влияющих факторов.
2. Недостаточны размеры обучающей выборки.
3. Возможно улучшить точность прогнозирования изменением представления данных, например, заменой натурального представления на относительное.

Выводы

В результате можно сделать выводы:

1. Точность решения задачи классификации с использованием данного метода достаточна для дальнейшего использования.
2. Необходим дополнительный поиск влияющих на прогнозируемую величину факторов.
3. Для повышения точности прогнозирования возможно необходимо изменить представление факторов в выборках.

Вопрос о развитии методики прогнозирования применительно к данному предмету исследования, таким образом, остается открытым. В данном случае развитие методики можно связать прежде всего с предварительной обработкой данных: удаление влияющих факторов,

преобразование представления данных для сокращения размера выборок и улучшения различимости ситуаций.

Литература

1. Хмелевой С.В., Скобцов Ю.А. Нейросетевой поход к прогнозированию цен на подержанные автомобили.// Наукові праці ДонНТУ серія: обчислювальна техніка та автоматизація. Вып. 74. –Донецк: изд-во «Лебідь», 2004г., С. 140 – 147.

2. Д.-Э. Бэстенс, В.-М. Ван Ден Берг, Д. Вуд «Нейронные сети и финансовые рынки. Принятие решений в торговых операциях». – М.: Научное издательство ТВП, 1997. 235с.

3. Proben1 – a set of Neural Network Benchmark Problems and Benchmarking Rules. - ftp.ira.uk.de/pub/neuron

4. Родионов П.Е. Краткосрочное прогнозирование котировок ОГВВЗ с использованием аппарата нейронных сетей. Сборник «Интеллектуальные технологии и системы» под ред. Ю.Н.Филипповича; Изд-во МГТУ им.Баумана, Москва, 1998г.

5. Медведев В.С., Потемкин В.Г. «Нейронные сети». – М.: ДИАЛОГ-МИФИ, 2002. – 496с.

6. Головко В.А. Нейронные сети: обучение, организация и применение. – М.: Радиотехника, 2001.- 256 с

7. Он-лайновый учебник // Статистический портал.- http:// www.statistica.ru.

8. Maxwell T, Giles C, Lee Y, Chen H. Nonlinear Dynamics of Artificial Neural Systems // Proceedings of the conference on neural networks for computing. – Washington (D.C.), 1986

Дата надходження до редакції 12.06.2005 р.